From Microbenchmarks to LLM Inference – An End-to-End Analysis on the Energy Efficiency of the Grace Hopper Superchip

Markus Velten, Christian von Elm, Lena Jurkschat, Gülçin Gedik, Sebastian Döbel, Daniel Hackenberg
ZIH / ScaDS.AI, CIDS, TU Dresden
Dresden, Germany

***Abstract*—Energy-efficient AI systems require not only efficient hardware, but also trustworthy power measurement and predictable power control. The Nvidia Grace Hopper Superchip (GH200) combines a 72-core Grace CPU with a Hopper H100/H200 GPU on a single board, sharing an adjustable power budget. We present an end-to-end energy-efficiency analysis covering internal power-sensor precision via microbenchmarks, configurable power-knob effects on HPL and HPL-MxP benchmarks, and energy-efficiency and total cost of ownership (TCO) for Llama-3.1-8B-Instruct inference. Internal sensors prove highly precise relative to an external calibrated power meter, supporting energy accounting in shared HPC or cloud environments. However, power-budget enforcement exhibits undocumented gaps. For selected benchmark and LLM regimes, moderate power caps improve energy efficiency. For the evaluated LLM workload, however, reduced limits do not improve TCO because throughput losses increase amortized hardware cost per token.**

***Index Terms*—Nvidia Grace Hopper, GH200, energy efficiency, power sensors, LLM inference, HPL, benchmarking**

I. Introduction

GPU-accelerated HPC systems dominate both the TOP500 and Green500 rankings. The Nvidia Grace Hopper Superchip leads the Green500 list and appears in multiple TOP500 Top 10 entries. Unlike conventional GPU servers that couple a CPU to a discrete GPU via PCIe, the GH200 places the Grace CPU and Hopper GPU on a single package connected by a cache-coherent NVLink interface at 450 GB/s per direction. This tight integration eliminates GPU memory-size and PCIe bandwidth bottlenecks while enabling a shared, user-configurable power budget ranging from 200 W to 1000 W at the module level.

Despite prior work on GH200 performance, energy efficiency, and LLM energy use, the combined effects of sensor precision, shared CPU/GPU power limits, workload phase behavior, and TCO remain underexplored. We address this gap with an end-to-end analysis from microbenchmarks to LLM inference. This is a summary of our full paper [1], which contains the complete methodology, additional results, and discussion.

II. System Architecture and Power Interfaces

The evaluated system is a QuantaGrid S74G-2U housing one GH200 480 GB module (see Fig. 1). Software includes PyTorch 2.6.0, vLLM 0.8.6, and CUDA 12.8. Nvidia exposes four hierarchical on-chip power domains through the hwmon sysfs interface (CPU, Grace, Module, System I/O) as well as module and GPU readings via NVML. Power limits for the module (200–1000 W, default 1000 W) and GPU (1–900 W, default 900 W) can be configured independently. An external ZES ZIMMER LMG95 calibrated power meter sampling at 20 Hz provides a ground-truth reference throughout our experiments.

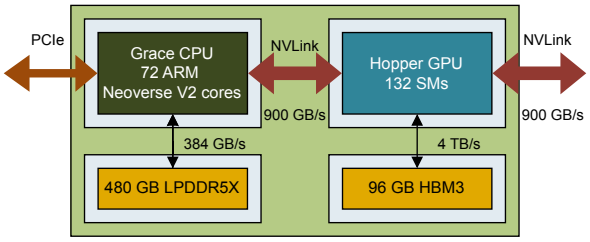


Fig. 1. Grace Hopper Superchip architecture. The Grace CPU (72 Arm Neoverse V2 cores, 480 GB LPDDR5X at 384 GB/s) and the Hopper H200 GPU (132 SMs, 96 GB HBM3 at 4 TB/s) share a module power budget and communicate via bidirectional NVLink at 900 GB/s.

III. Power Sensor Precision

We ported the roco2 microbenchmark suite to ARM64 SVE2, sweeping CPU-set sizes (8–72 cores in steps of 8), 30 frequency levels, and multiple C-state configurations. The resulting 1400+ operating points cover the full module power range and expose any sensor modeling artifacts. Fig. 2 confirms that both NVML and hwmon module power track the external meter with near-unity R^2 coefficients, unlike contemporary x86 processors. A systematic 9–13 W offset between the two internal interfaces suggests differing sensor post-processing, though both remain proportional to true consumption. Earlier anomalies reported by Yang et al. [2]—100 W hwmon noise swings and erroneous GPU power inclusion in nvidia-smi—appear resolved in our software stack (driver 570, CUDA 12.8), likely through firmware or driver updates.

IV. Power Knobs and HPL Benchmarks

We characterize the shared power budget using HPL (FP64) and HPL-MxP (FP8/FP16) from the Nvidia HPC-Benchmarks 25.04 container, sweeping all module and GPU power limit combinations on a 100 W grid (Fig. 3).

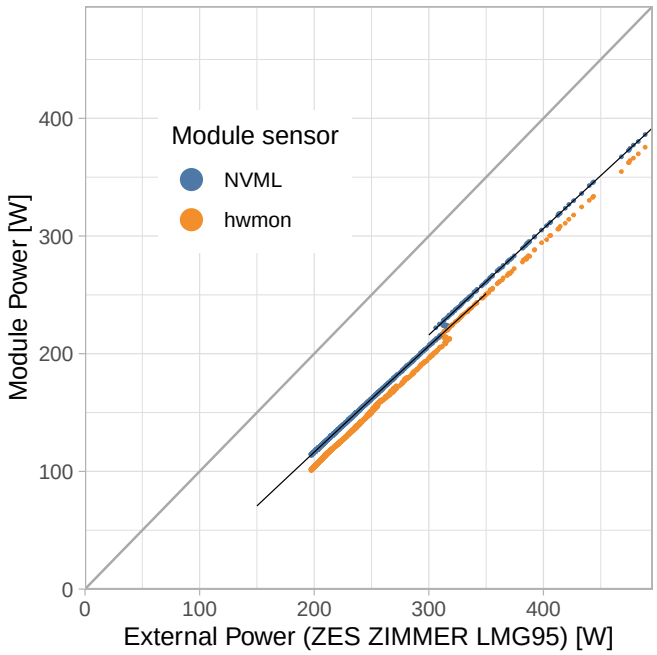


Fig. 2. roco2-based power sensor precision analysis. Internal NVML and hwmon module power versus calibrated external power. Black lines show quadratic fits to the lower (<225 W) and upper (≥225 W) ranges. Both NVML ($R^2 \approx 0.9998$–1.0) and hwmon ($R^2 \approx 0.999$–0.9997) track the external meter closely.

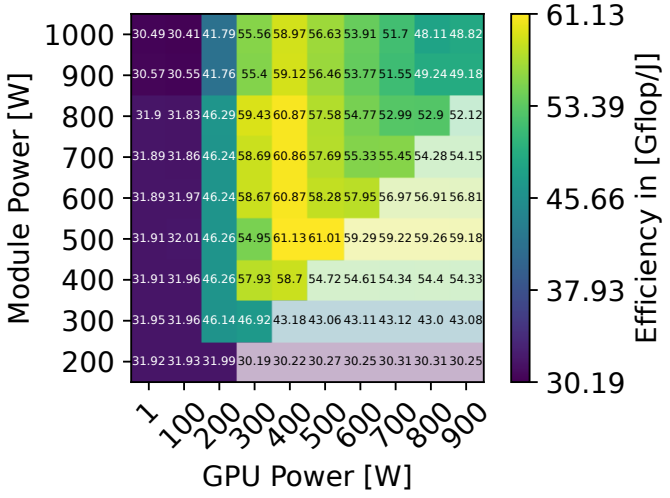


Fig. 3. HPL energy efficiency [Gflop/J] on the Hopper GPU (GPU-memory-only mode) across all module and GPU power limit combinations. Peak efficiency of 64.36 Gflop/J is achieved at 400 W module and 400 W GPU. De-saturated cells indicate GPU power limit exceeding module power limit.

HPL (FP64): Peak efficiency reaches 64.36 Gflop/J at 400 W module / 400 W GPU—comparable to the JUPITER Booster exascale system (63.3 Gflop/J, ranked 14th in Nov. 2025 Green500). The GPU power limit can be set above the module power limit; Nvidia does not document this behavior, and the GPU simply saturates at the module budget. At the lowest possible limits (200 W module / 1 W GPU),

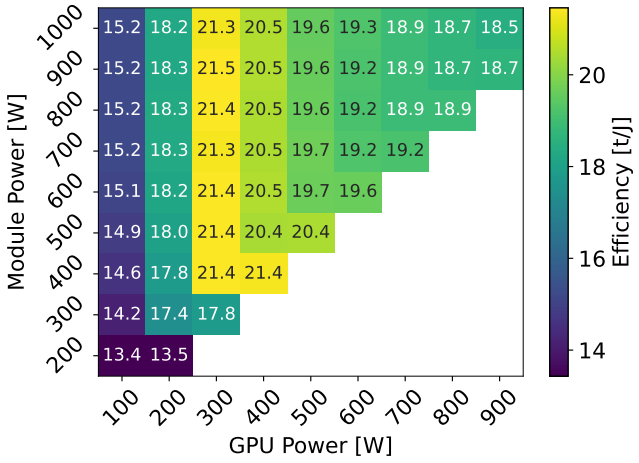


Fig. 4. LLM text-generation energy efficiency [tokens/J] for Llama-3.1-8B-Instruct served by vLLM across batch sizes and input lengths, at default power limits. Decoding-dominated configurations (large output lengths, large batch sizes) yield the highest efficiency and dominate operating costs.

measured module power still reaches ≈380 W (NVML), showing that the CPU's ≈325 W draw is not curtailed by the module power limit—a potentially hazardous undocumented behavior for operators relying on power capping.

HPL-MxP (FP8): The most efficient configuration (700 W module / 600 W GPU) achieves 816 Gflop/J at 446 Tflop/s, a 26% efficiency gain over the default-limit setting for a 9% performance reduction—a 12.7× improvement over FP64 HPL. FP8 and FP16 show structured efficiency saturation curves, whereas the CPU-only FP32 mode is insensitive to power limits (≲1% variation), consistent with the observed inability to cap CPU power independently.

V. LLM Inference: Energy Efficiency and TCO

We serve *Llama-3.1-8B-Instruct* with vLLM 0.8.6 and measure energy efficiency (tokens/J) and throughput (tokens/s) across input lengths (128–8192 tokens), output lengths (128–2048 tokens), and batch sizes (1–128), sweeping module and GPU power limits (Fig. 4).

Prefill vs. decoding: LLM prefill benefits from reduced power limits only with latency-bound configurations (e.g. small batch sizes and sequence length); power capping otherwise causes significant throughput reductions for prefill-heavy, compute-bound workloads. Decoding-heavy workloads (e.g. ≥512 output tokens) favor power limits, yet still favor maximum throughput once TCO is considered, so power reduction does not reduce cost.

TCO analysis: OPEX tracks energy efficiency and peaks at moderate GPU power limits (≈300–500 W). CAPEX, however, is throughput-driven: any reduction in generated tokens per second inflates the amortized hardware cost per token. The combined TCO therefore peaks with high-throughput configurations. For output lengths up to 2048 tokens, power limits

provide no TCO benefit; commercial API costs (≈52.36 Mt/€) are in the same order of magnitude as our on-premise TCO results, confirming that the GH200 is cost-competitive but does not gain from aggressive power capping in typical LLM serving scenarios.

VI. Conclusions

This work presents the first comprehensive end-to-end energy-efficiency study of the Nvidia Grace Hopper Superchip spanning microbenchmarks, power-knob exploration, and production-grade LLM inference. Key findings are: (1) internal power sensors are precise and suitable for workload-level energy accounting; (2) module power limits are not reliably enforced for CPU-heavy workloads, posing operational risks; (3) FP8 workloads exhibit strong efficiency gains from moderate power capping, while LLM decoding does not; and (4) for realistic LLM inference, maximizing throughput is the dominant lever for cost efficiency, not power reduction. Future work will extend the sensor precision analysis to other architectures and explore targeted per-domain power management.

Acknowledgment

The authors acknowledge funding by the Bundesministerium für Bildung und Forschung via the EECliPs project (grant no. 16ME0602) and support of the NHR Graduate School of the NHR-Verein e.V.

References

[1] Markus Velten, Christian von Elm, Lena Jurkschat, Gülçin Gedik, Sebastian Döbel, and Daniel Hackenberg. From microbenchmarks to llm inference an end-to-end analysis on the energy efficiency of the grace hopper superchip. In *2026 25th International Symposium on Parallel and Distributed Computing (ISPDC)*, pages 121–129, 2026.

[2] Zeyu Yang, Karel Adamek, and Wesley Armour. Accurate and Convenient Energy Measurements for GPUs: A Detailed Study of NVIDIA GPU's Built-In Power Sensor. In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2024.